An effective Named Entity similarity metric for use with syntax independent data

Croft, D, Brown, S & Coupland, S

Original citation & hyperlink:
Croft, D, Brown, S & Coupland, S 2016, 'An effective Named Entity similarity metric for use with syntax independent data' *Digital Scholarship in the Humanities*, vol 32, no. 4, pp. 779-787
https://dx.doi.org/10.1093/llc/fqw035

DOI 10.1093/llc/fqw035
ISSN 2055-7671
ESSN 2055-768X

Publisher: Oxford University Press

OXFORD UNIVERSITY PRESS | Digital Scholarship in the Humanities

An effective Named Entity similarity metric for use with syntax independent data

Journal:	*Digital Scholarship in the Humanities*
Manuscript ID:	LLC-2015-0018
Manuscript Type:	Full Paper
Date Submitted by the Author:	30-Mar-2015
Complete List of Authors:	Croft, David; De Montfort University, Faculty of Art, Design & Humanities Brown, Stephen; De Montfort University, Faculty of Art, Design & Humanities Coupland, Simon; De Montfort University, Faculty of Technology
Keywords:	Similarity, Named entity, Jaro, Jaro-Winkler, NESim, GLAM, Heritage, FuzzyPhoto

	benjamin	frances	johnston
b	0.71	0.00	0.00
francis	0.49	0.94	0.51
johnston	0.47	0.51	0.00
miss	0.00	0.00	1.00

Jaro-Winkler similarity matrix.
21x7mm (300 x 300 DPI)

	benjamin		*frances*		*johnston*
b	0.71	francis	0.94	johnston	1.00
francis	0.49	johnston	0.51	francis	0.51
johnston	0.47	miss	0.00	miss	0.00
miss	0.00	b	0.00	b	0.00

Ordered Jaro-Winkler similarity matrix.
31x13mm (300 x 300 DPI)

	john		j		doe
john	1.00	john	0.78	doe	1.00
doe	0.53	doe	0.00	john	0.53
smith	0.00	smith	0.00	smith	0.00

Match collision example.
17x3mm (300 x 300 DPI)

	john		j		doe
john	1.00	john	0.78	doe	1.00
doe	0.53	doe	0.00	john	0.53
smith	0.00	smith	0.00	smith	0.00

	john		j		doe
john	1.00	doe	0.00	doe	1.00
doe	0.53	smith	0.00	john	0.53
smith	0.00			smith	0.00

	john		j		doe
john	1.00	smith	0.00	doe	1.00
doe	0.53			john	0.53
smith	0.00			smith	0.00

Match collision resolution example.
59x46mm (300 x 300 DPI)

	benjamin b	frances francis	johnston johnston
Jaro-Winkler	0.71	0.94	1.00
Length	9	14	16
Weight	0.23	0.36	0.41
Combined	0.16	0.34	0.41
Result	0.91		

Combining element pair values.
30x13mm (300 x 300 DPI)

Thresh	Person			NESim		
	True pos	False pos	F-score	True pos	False pos	F-score
0.01	2792107	302798	0.690	4995922	4872474	0.672
0.10	2792107	302798	0.690	4995902	4871931	0.672
0.20	2778772	295806	0.688	4993946	4783092	0.676
0.30	2777763	291702	0.688	4978056	4185009	0.703
0.40	2726931	272477	0.682	4939370	3115295	0.757
0.50	2667683	192543	0.679	4853975	1752345	0.836
0.60	2592929	92813	0.675	4667930	486640	0.919
0.66	2456993	57263	0.654	4507634	149746	0.934
0.70	2221751	37656	0.612	4340600	55103	0.924
0.80	1545205	4106	0.472	3283597	2114	0.793
0.90	1047710	940	0.346	1160086	68	0.377
1.00	785085	508	0.271	79064	1	0.031

Subsection of test results.
44x21mm (300 x 300 DPI)

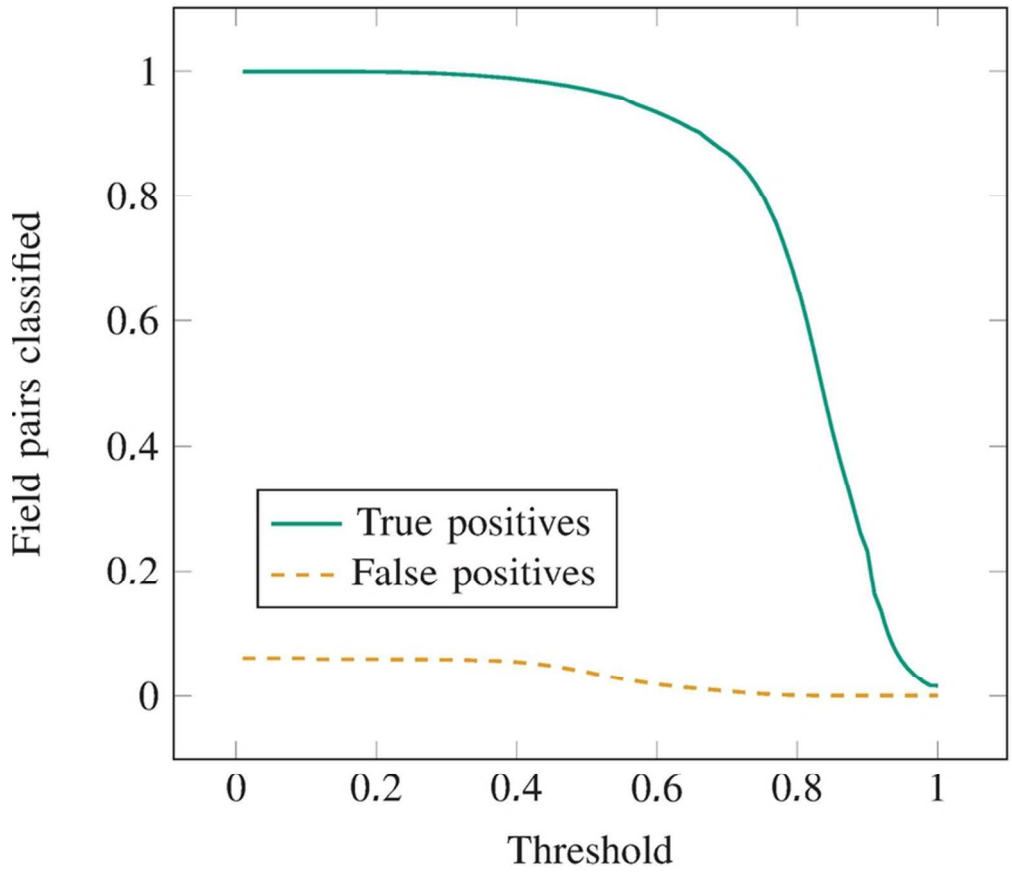
	Person	NESim
Total negative cases	$1 \cdot 10^7$	
Total positive cases	$1 \cdot 10^7$	
Area under the curve	0.81	0.46
Std error	0.0001	0.0002
Area difference	0.35	
Std error difference	0.0002	
Z	1538.3791	
P, non-directional	< 0.000001	
P, directional	< 0.000001	

Significance of difference between the two ROC curves.
32x19mm (300 x 300 DPI)

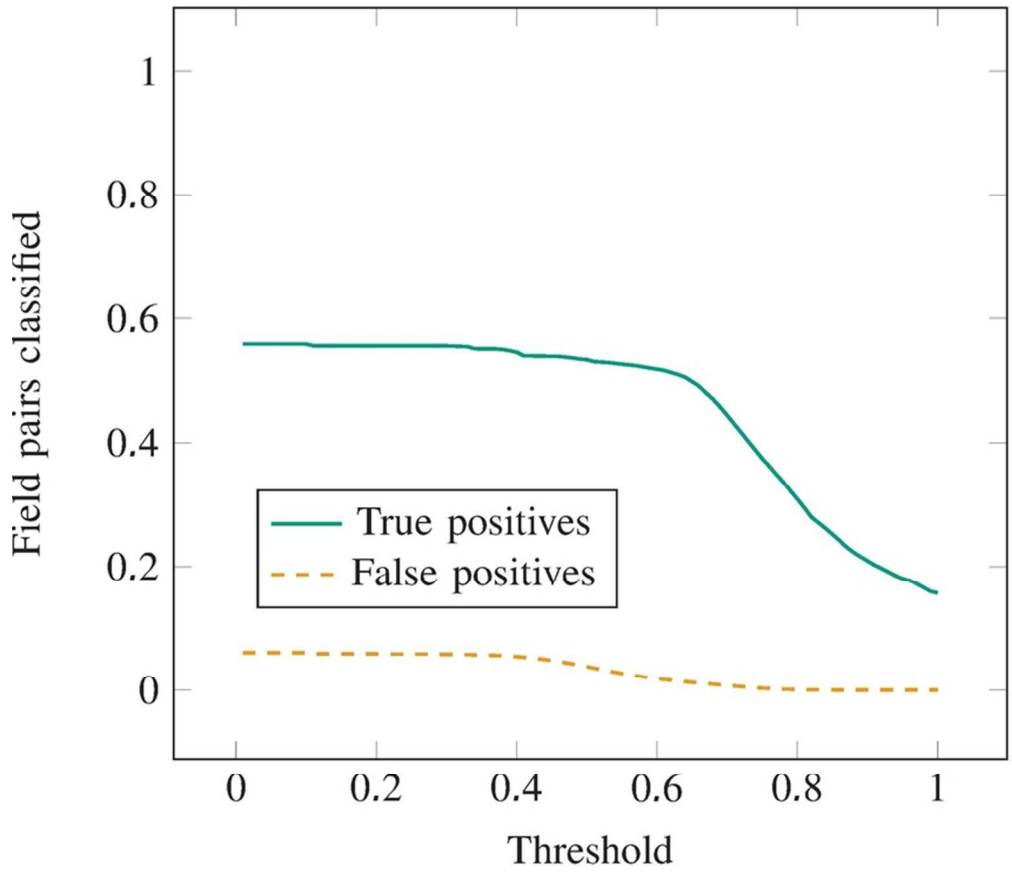
$$\frac{1}{2}\left(s + s\frac{2 \cdot (|A| \vee |B|)}{|A| + |B|}\right)$$

$$\frac{1}{2}\left(0.91 + 0.91\frac{2 \cdot (3 \vee 4)}{3 + 4}\right) = 0.845$$

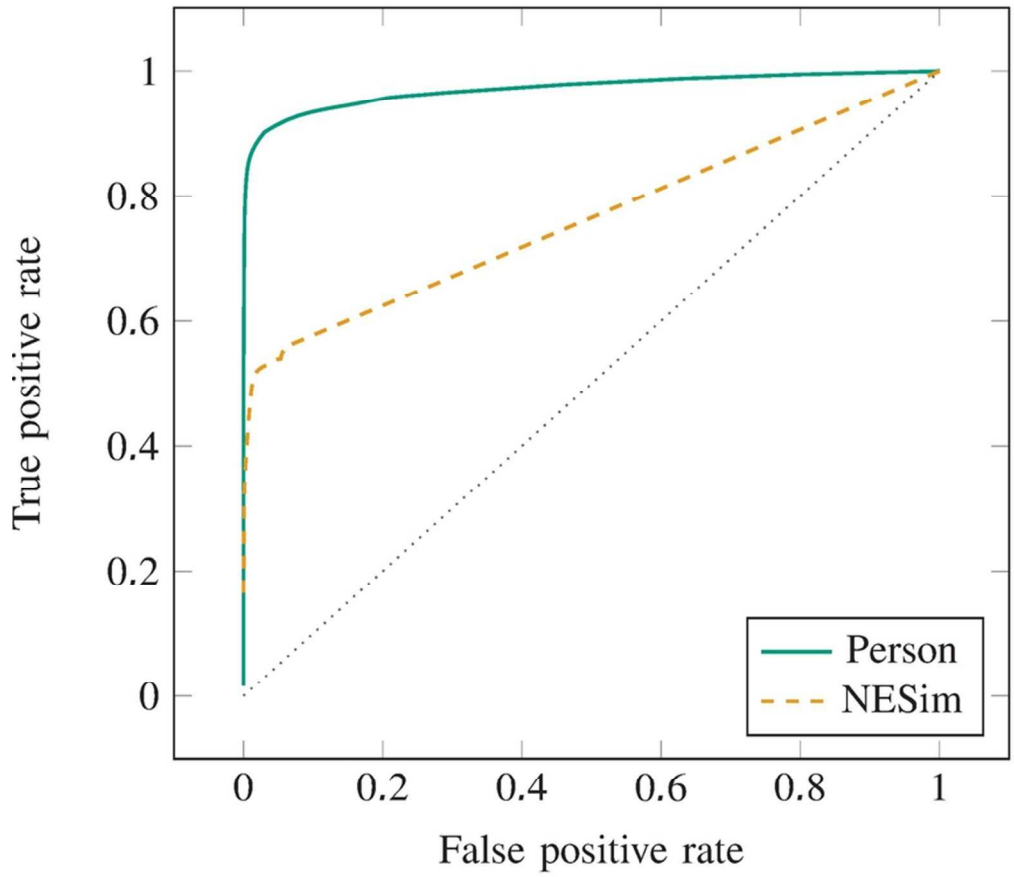
21x9mm (600 x 600 DPI)



True/False positive rates for person metric versus threshold value (excluding 0.0).
71x62mm (300 x 300 DPI)



True/False positive counts for NESim versus threshold value (excluding 0.0).
71x62mm (300 x 300 DPI)



ROC curve comparison of NESim and our person metric's performance.
72x63mm (300 x 300 DPI)

An effective Named Entity similarity metric for use with syntax independent data

David Croft, Stephen Brown and Simon Coupland

David Croft and Stephen Brown are with Knowledge Media and Design, De Montfort University, Leicester, LE1 9BH, United Kingdom (email: {dcroft, sbrown}@dmu.ac.uk).

Simon Coupland is with the Centre for Computational Intelligence, De Montfort University, Leicester, LE1 9BH, United Kingdom (email: simonc@dmu.ac.uk).

Correspondence

David Croft, 4 Cranedown, Lewes, East Sussex, BN7 3NA

dscroft@gmail.com

Abstract

This paper describes and demonstrates a names entity similarity metric developed for, and currently in use by, the FuzzyPhoto project. The presented metric is effective at comparing named entity data in and across syntax less data schemas such as are often encounter in GLAM collections. The efficiency of the approach was compared to an existing named entity similarity metric and is shown to be a significant improvement when comparing messy named entity data.

1 Introduction

FuzzyPhoto was a 2 year Arts and Humanities Research Council (AHRC) funded project that successfully developed and deployed a Fuzzy Inference System (FIS) based data mining system to identify co-referent records in the photographic collections of multiple Gallery, Library, Archive and Museum (GLAM) institutions.

The data mining system produced uses four fields extracted from the full collection records to try and identify occurrences of the same photograph across multiple institutions. The four fields are:

1. Title - A short textual description of the photographs contents.
2. Person - Typically the name of the photographer.
3. Process - The chemical and mechanical process/es used in creating the image.
4. Date - The creation date of the photograph.

By combining these four fields, the FuzzyPhoto project demonstrates that it was possible to identify matches across and within GLAM collections despite imprecision and uncertainty of the information held and the differing schemas in use across the sector (Brown, Coupland and Croft 2013). A secondary outcome, however, was the creation of multiple similarity metrics each tuned to the specific challenges of GLAM collection data and the difficulties of a specific field. This paper presents one of these metrics, that deals with person data.

2 Person field

The person name field is obviously important for identifying similarities between records in different archives. In FuzzyPhoto, the contents of the person field for each pair of records being compared are processed to identify if the same entity is described in both. While the person field typically contains the name of the person that took the photograph, the names of photography studios are also frequent. Named entity comparison and identification is a well

established problem in a variety of domains. The comparison of named entity in GLAM records is, however, particularly challenging due to a combination of factors. These are:

1. Typographical errors and variations - These include everything from simple spelling mistakes made when the name information was first recorded**[NOTE 1]** to transcoding errors made when information was digitised.
2. Extraneous information - Some GLAM institutions store more than just name information in the person fields of their databases. The most common offenders are the birth and death dates of the individual named but addresses and job titles also make appearances. Whilst this information can be useful for researchers, its inclusion in name fields is undesirable and would ideally have been stored in separate fields.
3. Short forms - Within GLAM collection records the problem generally manifests as comparing full names to names containing initials**[NOTE 2]**. Outside of GLAM collections this problem can also be extended to cover the difficulties in comparing full names to nicknames/variations, “Robert” is an obvious example having multiple valid variations including Bob, Bert, Rob and Robby.
4. Name order - Unlike commercial customer databases where the practise is to store individual name elements separately (i.e. separate forename and surname fields), GLAM collection typically have a single name field containing all of the name data. As names can can be written in a number of different ways**[NOTE 3]** records will ideally conform to a standard syntax.

These are known problems in the areas of name and textual similarity. Typographical variations are a problem faced by any name comparison system and as such there are a number of existing approaches which allow for name matching despite said variations. These include phonetic approaches such as Soundex (Odell and Russell 1918) and Metaphone (Philips 2000), edit distance approaches such as Damerau-Levenshtein (Damerau 1964) and others such as Jaro (Jaro 1989) and

Jaro-Winkler (Winkler 1990).

Methods for recognising and extracting named entities are also well established with multiple name identification techniques already available (Nadeau and Sekine 2007). Named entities and extraneous information being stored together is not, therefore, an insurmountable problem. However, the use of most named entity recognition techniques is excessive in this scenario which is an inversion of the normal situation. In this case the presence of a name is already known and it is only a minimal quantity of non-name data that needs to be removed.

Comparison of short and long form names can be addressed using many of the same techniques used to handle typographical variations, given that most short form names are just truncated versions of the long form.

The primary problem for name comparison in GLAM records is the name order. As stated previously, if all of the known name information is to be stored in a single field then those fields will ideally conform to a known syntax. This is not the case. GLAM collection records are not wholly consistent in the syntaxes they use to store name information. This applies not just between the collections of differing institutions, but
frequently within individual institutions and collections.

In situations such as those of the FuzzyPhoto project, where the name information from multiple distinct collections is collected and compared, the use of differing name orders between differing collections would be an irritating but easily solved problem. The name information could be converted into a single standard representation during the record acquisition. However, the widespread use of syntax independent metadata schemas within the GLAM sector means that the name order used within individual collections varies from record to record. Even name order is not a problem for some techniques (i.e. Named Entity Similarity (NESim)**[NOTE 4]**).

In this paper we present our method for unformatted entity name comparison. Our approach is able to perform named entity comparison in a computationally efficient manner and significantly

outperforms an existing technique in this situation.

This paper contains a description of our metric, worked examples of the metric and performance comparison of our metric against that of NESim. We demonstrate that our method's performance is significantly better than that of NESim when comparing unstructured entity names as encountered in GLAM collection records.

3 Person metric

In this section we describe our approach in detail and include a worked example in order to clarify certain sections.

In the worked example, the field values used are "johnston, frances benjamin, 1864-1952"**[NOTE 5]** and "Miss Francis B. Johnston"**[NOTE 6]**. These values demonstrates the ability of our approach to handle differences in element ordering, initials and additional information.

3.1 Tokenisation/filtering

The first stage is tokenisation and filtering of the raw data. The raw text is converted to lower case and split into separate elements at the word boundaries. Word boundaries are considered to be anywhere a punctuation character**[NOTE 7]** is found. Non-alphabetic characters are removed. For the worked example this produces the two vectors seen below:

- A = ['benjamin', 'frances', 'johnston']
- B = ['b', 'francis', 'johnston', 'miss']

As our research has, so far, been focused entirely on collections from Western Europe, North America and other English speaking counties, our approach has only been designed to work with the Latin character set. This would restrict our approach the Germanic (e.g. English and German) and Romance languages (e.g. French, Spanish etc). At present the C++ Jaro-Winkler implementation (see section 3.2) we use only supports American Standard Code for Information

Interchange (ASCII) coded strings, however Unicode supporting implementations do exist in other programming languages. We are, therefore, hopeful that our approach can be expanded to handle other encodings in the near future, for the moment non-ASCII characters should be converted to their base forms (e.g. ò ó ô õ ō ȯ ö ọ ŏ ǒ → o) before being processed by the metric and it does not work for languages such as Russian, Japanese or Arabic where no ASCII compatible base form of the character sets exist.

3.2 Element similarity

The second stage is the generation of a complete similarity matrix for elements of the two vectors being compared. This has to potential to be computationally expensive for vectors with a large number of elements, however the average number of elements is only 0.34**[NOTE 8]**. The matrix sizes we are producing here are, therefore, low. The resulting matrix for the worked example can be seen in table 1.

Jaro-Winkler was used for the individual elements comparison over other techniques (specifically Jaro) as it applies additional significance to the start of the terms being compared. As mentioned in section 2, one problem with name comparisons are initials and alternate short forms of full names. Obviously initials will be based on the first letter of a full name but short forms are also predominately based on the start of a full name rather than the middle and end (e.g. Dave from David, Matt from Matthew) although exceptions exist (e.g. Beth from Elizabeth, Dick from Richard). As handling these forms would likely require a database nicknames which would increase the complexity and processing time of our approach these exceptions are ignored. **[TABLE 1 HERE – table1.eps]**

3.3 Pair selection

Step 3 is selecting the element pairs from the matrix. Our approach attempts to find the best overall (i.e. the configuration of non-overlapping element matches that produces the highest combined

Jaro-Winkler similarity value). Whilst an exhaustive search of all of the possible combinations (brute force) would guarantee that the optimum solution was found, this results in excessive computational requirements for the approach and is rarely necessary. Instead the combination of element pairs is selected heuristically.

Element pairs are selected by ordering the Jaro-Winkler similarity values. For each element of A the similarity values against *B* are ordered from highest to lowest, see table 2.

The best match between each element of *A* to an element from *B* is then selected as the 1st pair in each ordered column. In the case of the worked example the best pairs are 'benjamin' ↔ 'b' = 0.71, 'frances' ↔ 'francis' = 0.94 and 'johnston' ↔ 'johnston' = 1.00. **[TABLE 2 HERE – table2.eps]**

Although in our earlier example every element of A matched against a different element of B in the order similarity matrix this will not always be the case**[NOTE 9]**. Under our approach two or more elements in one vector are not allowed to match against the same element in the other.

If a collision is detected then at least one of the selected pairs must be replaced. The pair with the lowest value should be changed. In cases where multiple matches have the same value, the match which will produce the smallest change should be chosen. If multiple matches will produce the same change, select the first one.

The following section demonstrates a collision situation and the pair alterations required. For this example, the two vectors in this case are *C* = ['john', 'j', 'doe'] and *D* = ['john', 'smith', 'doe'].

The resulting ordered Jaro-Winkler similarity matrix is shown in table 3. As that table shows, there is a collision between the 'john' and 'j' elements in *C* where both have matched to the 'john' element in *D*. **[TABLE 3 HERE – table3.eps]**

In this case the correct action is to change the 'john' ↔ 'j' match instead of the 'john' ↔ 'john' match as this has a similarity of 0.78 as opposed to 1.00. Unfortunately making said change produces a new collision and so the process must repeat again, the full list of changes can be seen in table 4. **[TABLE 4 HERE – table4.eps]**

For each element of A to match against a different element of B, $|A| \leq |B|$ must be true. This is easily achieved by simply assigning the shortest vector to be A, however in cases where $|A| = |B|$ then the element selection should be conducted twice with Jaro-Winkler similarity matrix transposed between iterations.

3.4 Match weighting

The Jaro-Winkler values for the element matches are then weighted according to the combined length of each pair as a proportion of combined length of all the combined pairs. This weighting means that matches between two initials or matches between an initial and a full name are considered to be less significant than matches between longer elements. Although two initials could be identical it does not mean that the full names they represent are the same, our weighting approach allows the match between initials to contribute to the overall match value but also recognises its inherent uncertainty. This effect of this weighting is shown in table 5. **[TABLE 5 HERE – table5.eps]**

3.5 Overall weighting

Finally the overall similarity value is weighted according to the proportion of the elements from $A + B$ that were paired. If, for example, we were to compare vector A (from the worked example) against another vector E = ['benjamin'], then under the approach described so far that would produce an overall similarity value of 1.0. Therefore in order to take into account the number of elements actually compared and so rank $A \leftrightarrow B < A \leftrightarrow E$, the similarity value is modified as shown in equation 1 where s is the unmodified similarity value. **[EQUATION HERE – equation.tif]**

4 Testing

4.1 Dataset

In order to measure the performance of our name matching approach, we attempted to identify co-referent entity names in a pre-labelled testing dataset. We made use of a the JRC-Names (version 1) dataset (Steinberger, Pouliquen, Kabadjov and der Goot 2013) produced by the Joint Research Centre (JRC) of the Institute for the Protection and Security of the Citizen (IPSC). JRC-Names is a list of 573,141 entries describing variations on 268,521 distinct names. Included in JRC-Names are a number of entity names (e.g. places, companies). Although our approach was designed for person names, it also functions effectively with non-person names and so these entries were left in the dataset.

In order to better represent the formats and information present in actual GLAM records, the JRC-Names dataset was expanded to include poorer quality versions of the existing entries. The changes made to the original entries include, adding title information, changing the name order, removing middle names and shortening full names to just their initials.

For example, the entry “Ira Lee Sorkin” in the original JRC-Names dataset was modified and expanded to include such variations as “Ms I L Sorkin”, “Sorkin, I” and “Mrs Ira L Sorkin” amongst others. In total the expanded dataset contained 179,490 entries. Whilst the expanded dataset is not a perfect model of the information found in GLAM records, specifically it is lacking examples of extraneous information. However, it does include instances of all the other problems discussed in Section 2. Pairs of elements were randomly selected from the expanded dataset to generate 1•10□ test cases split evenly between positive and negative cases.

4.2 Experimentation

There are two factors which must be considering the effectiveness of our metric for identifying co-referent entity names, the recall and precision of our approach. These will of course be affected by the value used as a threshold to distinguish between co and non-referent entities.

Although our usage of this metric in the FuzzyPhoto is as one input to a Mamdani style FIS, we

have chosen to use it as part of a simple threshold approach here in order to demonstrate its effectiveness. The 1•10□ test cases were run through both our person metric and NESim. NESim was chosen as a test candidate as it has proven to be effective in addressing the most issues specified in Section 2. The only issue NESim has proven ineffective in addressing is that of extraneous information. Examples of that issue were, therefore, not included in the testing data.

The true and false positive rates of the two approaches can be seen in figures 1 and 2. As these figures show, the true positive rate of our metric is significantly higher than that of NESim. In order to compare the relative performances of the two approach we utilise Receiver Operating Characteristic (ROC) curves. These can be seen in Fig.3, and clearly show that our person metric performs better than NESim. Optimum performance for both approaches, as measured by their F-scores, is seen with at a threshold value of 0.66. At this threshold value our approach classifies 93.4% of the test cases correctly. This is compared to peak performance 69.0% for NESim when using a threshold of $0 < t \leq 0.1$. The significance of this results was calculated using the approach described by Hanley and McNeil (1982) and produced a p value of < 0.000001 (see table 7).

[FIGURE 1 HERE – fig1.eps] [FIGURE 2 HERE – fig2.eps] [FIGURE 3 HERE – fig3.eps] [TABLE 6 HERE – table6.eps] [TABLE 7 HERE – table7.eps]

As both metrics were implemented in different languages, it was not possible to conduct a fair comparison of the processing throughputs of the two approaches.

5 Conclusion

As the ROC curves in Fig.3 and analysis in table 7 show, our person metric is highly effective at identifying co-referent entity names when compared to NESim. GLAM community collection records are an unusual and challenging comparison space and our results clearly demonstrate that a named entity similarity metric which is tuned to the specific challenges of the GLAM search space

can produce significantly better results than more established techniques.

Whilst our approach is currently limited to entity names written in Germanic or Romance languages due to our current implementation of Jaro-Winkler, we hope that it will be possible to expand it to cover a broader range of languages. Jaro variants are already known to be effective against certain Asian languages (Recchia and Louwerse 2013) and so a broader application of our metric is, in part, a matter of improving the software implementation.

Notes

1. Some GLAM collections attempt to digitise exactly the original information. This includes deliberately reproducing any errors which may exist in the original. This is also a problem for records created using Optical Character Recognition (OCR) methods.

2. “H. T. Malby” vs. “Malby, Henry Thomas” to give a real example.

3. For example, ‘forename initial surname’, ‘surname, forename initial’ and ‘initial surname, forename’. This is predominantly true for person names but can also apply to institutions or business names.

4. Under NESim “John Smith” vs “Smith John” produces a value of 1.0.

5. Copied exactly from a Library of Congress (LoC) records.

6. Copied from an Exhibitions of the Royal Photographic Society (ERPS) record but with a typographical error deliberately introduced, “Frances” → “Francis”.

7. i.e. commas, colons, semi-colons and spaces.

8. Based on an analysis of 342,797 records from 7 GLAM collections, the same records produced a maximum size of 20.

9. It is, however, rare.

10. Our approach was implemented in C++ whilst NESim was tested using a Python wrapper to feed the test cases to the Java implementation of NESim available on the CCG: Software page,

http://cogcomp.cs.illinois.edu/page/software/.

References

Brown, S., Coupland, S. and Croft, D. (2013), Where are the pictures? linking photographic records across collections using fuzzy logic., in N. Proctor and R. Cherry, eds, 'Museums and the Web Asia', 221–227.

Damerau, F. J. (1964), 'A technique for computer detection and correction of spelling errors', Commun. ACM 7, 171–176.

Hanley, J. A. and McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (roc) curve', Radiology 743, 29–36.

Jaro, M. (1989), 'Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida', Journal of the American Statistical Association 84(406), 414–420.

Nadeau, D. and Sekine, S. (2007), 'A survey of named entity recognition and classification', Lingvisticae Investigationes 30(1), 3–26.

Odell, M. and Russell, R. (1918), 'U.s. patent numbers 1,261,167'.

Philips, L. (2000), 'The double metaphone search algorithm', C/C++ users journal 18(6), 38–43.

Recchia, G. and Louwerse, M. (2013), A comparison of string similarity measures for toponym matching, in 'Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place', COMP '13, ACM, New York, NY, USA, pp. 54:54–54:61.

Steinberger, R., Pouliquen, B., Kabadjov, M. A. and der Goot, E. V. (2013), 'JRC-names: A freely available, highly multilingual named entity resource', arXiv preprint arXiv:1309.6162.

Winkler, W. E. (1990), String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage, in 'Proceedings of the Section on Survey Research', 354–359.

Figure Legends

1. True/False positive rates for person metric versus threshold value (excluding 0.0).

2, True/False positive counts for NESim versus threshold value (excluding 0.0).

3, ROC curve comparison of NESim and our person metric's performance.

Table Legends

1. Jaro-Winkler similarity matrix.

2. Ordered Jaro-Winkler similarity matrix.

3. Match collision example.

4. Match collision resolution example.

5. Combining element pair values.

6. Subsection of test results.

7. Significance of difference between the two ROC curves.